Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations

Evans, Luke; Slavov, Gancho; Rodgers-Melnick, Eli; Martin, Joel; Ranjan, Priya; Muchero, Wellington; Brunner, Amy M; Schackwitz, Wendy; Gunter, Lee E.; Chen, Jin-Gui; Tuskan, Gerald A.; DiFazio, Stephen P.

Published in:
Nature Genetics

DOI:
10.1038/ng.3075

Publication date:
2014

Citation for published version (APA):
Evans, L., Slavov, G., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., Brunner, A. M., Schackwitz, W., Gunter, L. E., Chen, J-G., Tuskan, G. A., & DiFazio, S. P. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nature Genetics*, *46*, 1089-1096. https://doi.org/10.1038/ng.3075

Population genomics of the model tree *Populus trichocarpa* identifies signatures of selection and adaptive trait associations

Luke M. Evans[1], Gancho T. Slavov[2], Eli Rodgers-Melnick[1], Joel Martin[3], Priya Ranjan[4], Wellington Muchero[4], Amy M. Brunner[5], Wendy Schackwitz[3], Lee Gunter[4], Jin-Gui Chen[4], Gerald A. Tuskan[3,4], Stephen P. DiFazio[1,6]

[1] Department of Biology, West Virginia University, Morgantown, WV 26506

[2] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, SY23 3EB, UK

[3] The Joint Genome Institute, Walnut Creek, CA 94598

[4] Plant Systems Biology Group, BioSciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

[5] Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA 24061

[6] Corresponding Author: spdifazio@mail.wvu.edu

ABTRACT:

Forest trees are dominant components of terrestrial ecosystems that have global ecological and economic importance. Despite distributions that span wide environmental gradients, many tree populations are locally adapted, and mechanisms underlying this adaptation are poorly understood. Here we use a combination of whole-genome selection scans and association analyses of 544 *Populus trichocarpa* trees to reveal genomic bases of adaptive variation across a wide latitudinal range. Three hundred ninety-seven genomic regions showed evidence of recent positive and/or divergent selection, and enrichment for associations with adaptive traits that also displayed patterns consistent with natural selection. These regions also provide unexpected insights into the evolutionary dynamics of duplicated genes and their roles in adaptive trait variation.

A suite of forces and factors, including mutation, recombination, selection, population history, and gene duplication influence patterns of intraspecific genetic variation. Distinguishing which factors have shaped sequence variation across a genome requires extensive whole-genome sequencing of multiple individuals, which has only recently become tractable[1]. Most large-scale whole-genome resequencing studies have focused on model and domesticated species[1–5]. However, large-scale genome sequencing of natural populations holds great promise for advancing our understanding of evolutionary biology, including identifying functional variation and the molecular bases of adaptation. Recent work in a number of species has identified genomic regions that show signatures of positive selection and infer that such regions contain loci that control

adaptive traits[4,6–8]. Relatively few studies, however, have combined genome-wide scans with phenotypic data to determine if computationally-identified selected regions influence adaptive phenotypic variation[5,9–13]. Genome-wide studies of large natural populations combined with phenotypic measurements are necessary to determine which factors shape patterns of genetic variation within species, and therefore enhance our understanding of adaptation.

With large geographic ranges spanning wide environmental gradients and a long history of research demonstrating local adaptation[14], forest trees are ideal for examining the processes shaping genetic variation in natural populations. Forest trees cover approximately 30% of terrestrial land area[15], provide direct feedback to global climate[15], and are often foundation species that organize entire biotic communities and biogeochemical systems[16,17]. Clearly, biotic and abiotic interactions have influenced population sizes and distributions of forest trees, leaving diagnostic signatures in the genomes of present-day populations[14,18,19]. A deeper understanding of the evolutionary and ecological forces that shaped these patterns will offer insights and options for ecosystem management, applied tree improvement, and accelerated domestication efforts[20].

Black cottonwood, *Populus trichocarpa* Torr. & Gray, is a dominant riparian tree that has become a model for the advancement of genomic-level insights in forest trees[21]. The sequencing of 16 *P. trichocarpa* genomes revealed widespread patterns of linkage disequilibrium (LD) and population structure[22] and extensive genecological studies have revealed a high degree of adaptive phenotypic variation in growth, vegetative phenology and physiological traits such as water use efficiency and photosynthesis[23–25], suggesting

that local adaptation is prevalent. To date, candidate gene association analyses have revealed loci with significant effects on phenotypic traits[26,27]. However, thus far there have been no publications describing whole-genome associations for adaptive traits in *P. trichocarpa*, and their relationship to signatures of selection in any forest tree species.

One of the salient features of the *P. trichocarpa* genome is a remarkably well-conserved whole-genome duplication that is shared by all members of the Salicaceae and near relatives: the Salicoid duplication[28,29]. Despite the extensive occurrence of segments of collinear paralogous genes, over two-thirds of the duplicate pairs have been lost since the duplication event and there are substantial functional biases in the remaining gene pairs, in particular, an overabundance of gene categories with large numbers of protein-protein interactions[30,31]. A major unexplored question is whether the fundamental, diagnostic differences in diversity between retained duplicate pairs and genes lacking paralogs from the Salicoid duplication (singletons) are connected to patterns of natural selection and adaptive phenotypic variation.

Here we report the whole-genome resequencing of a collection of 544 *P. trichocarpa* individuals, spanning much of the species' natural latitudinal range, that have been clonally replicated in three contrasting environments. We use this resource to detect signatures of recent selection across the *Populus* genome and on adaptive traits themselves. We also show that the signals of association with adaptive traits are stronger in positively selected regions. Finally, we demonstrate that Salicoid duplicate genes have distinctive patterns of adaptive variation that reveal the evolutionary effects of dosage constraints.

RESULTS

Polymorphism and population structure

From high-coverage whole-genome sequencing of 544 unrelated *P. trichocarpa* individuals (Fig. 1a, Supplementary Table 1) we collected over 3.2 Tbp of data that aligned to 394 Mbp of the *P. trichocarpa* genome. Approximately 87.5% of the 3.2 Tbp was accessible for analysis based on median sequencing depth across all samples (Supplementary Fig. 1). From these data, we detected 17,902,740 single nucleotide polymorphisms (SNPs).

Using this resource, there was a two-fold higher nucleotide diversity in intergenic sequence than in genic sequence, largely consistent with purifying selection (Table 1). Diversity was particularly low in coding sequence, where nonsynonymous diversity was one-third that of synonymous diversity. Most SNPs were rare (MAF≤0.01), particularly those predicted to have major effects (e.g., splice site mutations) (Table 1, Supplementary Fig. 2). We also identified 5,660 large (>100 bp) and 254,464 small (<50bp) insertion/deletion (INDEL) polymorphisms, which will be described in detail in a separate publication.

Based on principal components analysis (PCA) of all 17.9 million SNPs, we identified four major regional genetic groups corresponding to geographical origin (Fig. 1a). We also found genetic-geographical structure within regional groupings that clustered as separate subgroups within source locations (Fig. 1b). These data indicate that there is genome-wide genetic structure at both broad latitudinal and local spatial scales.

Phenotypic evidence of selection

We examined two different indicators of selection using phenotypic data from three clonally replicated plantations representing the center and southern extent of the extant range of *P. trichocarpa*. We found that quantitative differentiation (Q_{ST}) in height, spring bud flush, and fall bud set among source rivers was greater than genome-wide marker differentiation (F_{ST}) (Fig. 2a), suggestive of spatially divergent selection[32], as is commonly observed in forest trees[14,24,25]. Furthermore, at all three plantations, these same adaptive traits show correlations with multivariate climate variables (Fig 2b-d; Supplementary Fig. 3). Warmer climates (negative PC1) are associated with earlier bud flush and later bud set, strongly supporting the hypothesis that climate is a major determinant of adaptive genetic variation throughout the sampled range of *P. trichocarpa*[24,25].

Recent positive and divergent selection

We next attempted to relate the strong evidence of climate-driven, divergent selection on adaptive traits to genomic regions that also appear to be affected by natural selection. We examined five distinct metrics of natural selection using 1-kb windows across the genome. These metrics included allele frequency differentiation among subgroups (F_{ST}), allele frequency cline steepness across mean annual temperature and precipitation measurements (SPA[33]), extended haplotype homozygosity around alleles from rapid allele frequency increase (iHS[8]), and allele frequency clines with each of the first two climate principal components axes (bayenv[34], PC1 and PC2, respectively). From this data we classified the empirical top 1% of windows/regions as "selection outliers," i.e., regions with unusually strong polymorphism patterns consistent with recent

positive/divergent selection (Fig. 3, Supplementary Fig. 4 & 5, Supplementary Tables 2-6). Most of the selection outlier regions occurred uniquely among selection scan metrics, suggesting that each metric provides a distinct view of selection and that different selective forces are shaping these genomic regions (Fig. 3a). However, we found 397 regions in the top 1% for at least two of the selection scan metrics; we termed these regions "candidate selection regions" (CSRs) (Supplementary Table 7).

We tested whether the genes spanning or nearest to these CSRs (452 genes) and the selection outliers (1418, 1718, 1151, 257, and 312 genes for F_{ST}, SPA, iHS, bayenvPC1, and bayenvPC2, respectively) were overrepresented among annotation categories, gene families or genes with known involvement in several biological processes (Supplementary Tables 8-11, Fig. 3). Based on Fisher exact tests, certain functional categories were overrepresented, including GO annotations related to: response to stimuli, 1,3-β-glucan (callose) synthesis, and metabolic processes, as well as panther annotations for leucine-rich repeat receptor-like protein kinase and homeobox protein transcription factors (Supplementary Tables 8-10).

Despite some similarities, genes associated with the top 1% of each scan were generally overrepresented in unique categories (Fig. 3). For example, transcription factors (TFs) as a group were overrepresented among F_{ST} and SPA outliers; DELLA proteins (PF12041, gibberellin-interacting transcriptional regulators), among F_{ST} and bayenvPC2; and phytochromes (PF00360), genes involved in photoperiodic/circadian clock regulation, ATPase activity, and transmembrane movement (e.g., GO:0042626) were only overrepresented in F_{ST} (Supplementary Tables 8,9). Heat shock-related annotations were significantly overrepresented only in SPA (PTHR10015, PTHR11528), while proteins

induced by water stress or abscisic acid (PF02496) were overrepresented in bayenvPC2 and SPA outliers. 4-nitrophenylphosphatase, a hydrolase, was overrepresented among bayenvPC1 and weakly in F_{ST} (Supplementary Table 9). Class-III aminotransferases (PTHR11986, involved in abiotic stress[35]) were overrepresented most strongly in bayenvPC2 (Fig. 3).

Intriguingly, while moderate-effect SNPs were underrepresented among genic regions of all selection scan outliers, presumably due to purifying selection, SNPs with predicted high impacts were overrepresented among strong sweep loci implicated by the iHS scans (Supplementary Table 12), potentially because SNPs with major, presumably beneficial effects are more likely to be swept to high frequency. Because different selection processes (e.g., hard sweeps vs. subtle frequency shifts of standing variation) will influence diversity patterns differently, these five metrics reveal an assortment of potential selection pressures acting on *P. trichocarpa* through the largely non-overlapping regions identified in each

Adaptive trait associations in candidate selected regions

If climate is a major force driving the signatures of positive selection, we predict polymorphisms in these regions to be associated with climate-related adaptive traits. In particular, vegetative bud phenology should be a major determinant of fitness in these perennial populations, since timing of the onset and release of dormancy is largely shaped by photoperiod and temperature regimes[23,24]. Indeed, genes related to photoperiod, drought, and stress response were overrepresented among the selection outliers (Supplementary Table 11). To more directly test this hypothesis, we performed a

genome-wide association study (GWAS) with spring bud flush, fall bud set, and tree height measured at the three test sites, accounting for population stratification and background genetic effects in a mixed model framework for both univariate[36] and multivariate traits[37] (Fig. 1b, Supplementary Tables 1 & 13, Supplementary Fig. 6-10). More specifically, we found that those regions in the top 1% of scans had stronger adaptive trait association signals at all three test sites than expected by chance (i.e., the observed mean association signal was stronger than randomly resampled windows, controlling for gene density; all $p<0.00005$; Fig. 4, Supplementary Fig. 11). This was the case for all scans, including those based on spatial variation in allele frequency (e.g., F_{ST}, bayenv) as well as those based on long haplotypes (iHS). This correspondence is therefore unlikely to be artifactual, supporting the hypothesis that these outlier regions are partly driven by selection on adaptive traits.

We found strong associations for both univariate analyses as well as the multi-trait GWAS for each trait among test sites (Supplementary Table 13). Though some of the strongest univariate associations were also identified in the multiple-plantation GWAS, many associations were non-overlapping, perhaps due to the strong environmental differences among the locations, which ranged from cool and wet (Clatskanie, OR) to hot and dry (Placerville, CA). Strikingly few individual height-associated SNPs overlapped in comparisons between the Placerville, CA plantation and the other two sites.

Dormancy-related candidate genes in the selection and GWAS regions

A number of dormancy-related genes were near the strongest GWAS and selection signals. A region on chromosome 10, characterized by high LD, was one of the CSRs and was associated with bud flush ($p=5.19x10^{-6}$, Fig. 5). The strongest selection signal occurred near Potri.010G079600, a DNA-damage repair protein, and a number of lipid biosynthesis transferases. A strong bud set association also occurred near this region (Clatskanie and Corvallis, Supplementary Fig. 12). The strongest association signal ($p=5.69x10^{-7}$), within 15 kb of a CSR, was just downstream of the coding region of Potri.010G076100, a ureidoglycolate amidohydrolase (UAH) whose leaf and root expression is down-regulated with short days[38]. Ureides are transportable intermediates of purine catabolism, and by catalyzing the final step in ureide catabolism, UAH plays a role in the remobilization of nitrogen[39]. The ureide allantoin is also known to influence ABA metabolism and promotes abiotic stress tolerance in Arabidopsis[39]. However, to our knowledge, ureides and UAH have not previously been implicated as having important roles in seasonal N cycling or cold tolerance in *Populus*.

Among the photoperiodic and dormancy genes we found an F_{ST} outlier, Potri.010G179700 (*FT2*), which influences growth cessation in *Populus*[40]. This gene had an intronic SNP strongly associated with bud set and height ($p<0.00015$, Supplementary Table 13) and was near strong SPA and bayenv outliers. A second gene, Potri.008G117700 (similar to *PFT1*), occurred as an F_{ST} outlier region and was within 5 kb of several multi-trait association signals ($p=7.17x10^{-5}$). Arabidopsis *PFT1* is hypothesized to influence both defense and phytochrome B-mediated FT regulation[41].

Among the strongest bud flush associations ($p=2.72x10^{-14}$) was a nonsynonymous mutation in a 4-NITROPHENYLPHOSPHATASE locus, Potri.008G077400 (Clatskanie and Corvallis, Fig. 6). This mutation is in high LD with many other significantly associated SNPs in the surrounding 40 kb, including Potri.008G076800, (*FAR1* transcription factor) and Potri.008G077300 (UDP-galactose transporter), and is in an F_{ST} and bayenvPC1 outlier region. In this same region there is a bud flush association signal in all three test sites ($p=2.01x10^{-7}$ - $1.08x10^{-5}$) within Potri.008G077700 (*FT1*), a gene previously implicated in *Populus* dormancy cycling[42]. However, it appears to be an unlinked ($r^2=0.14$), separate association signal from that in Potri.008G077400.

In summary, we have detected genomic regions with patterns of diversity that are consistent with divergent and/or recent positive selection on a range of traits, and particularly on climate-related phenological and growth patterns. While our selection scans and GWAS analyses identified genes previously known to influence adaptive traits, they have also identified many loci of unknown function, which would not have been considered in any *a priori* candidate gene approach. Furthermore, the results and discussion presented above focus primarily on vegetative phenology, but many other traits are likely to be involved in determining fitness in these highly variable environments. In fact, the CSRs contained genes that have been implicated in controlling numerous other adaptive characteristics, including temperature stress tolerance, ion uptake and homeostasis, insect and pathogen defense, and reproduction. These are discussed in more detail in a Supplementary Note.

Duplication and Network Connectedness

We tested whether genes associated with selection outliers were over- or under-represented among the 7,906 identified gene pairs resulting from the Salicoid whole-genome duplication[29,31] (hereafter referred to as "Salicoid duplicates"), vs. genes that occur as singletons (Table 2). These analyses suggest that recent positive selective sweeps (indicated by iHS) are less likely for retained Salicoid duplicates than for singleton genes, but when one occurs, the sweep tends to occur for both duplicates. We also found that genes nearest to the individual F_{ST}, SPA, and iHS outliers had more predicted protein-protein interactions (PPI) than genes in the rest of the genome (Supplementary Fig. 13; $p \leq 0.05$). Furthermore, PPI were negatively correlated with nucleotide substitutions (π_T, π_S, and $\pi_{Nonsynymous}/\pi_{Synonymous}$ ratio; $r < -0.06$, $p < 0.0001$). These results suggest that patterns of selection (both purifying and positive) are influenced by genomic context, including past whole-genome duplication events and gene or protein-protein interactions. We discuss these analyses further in the Supplementary Note.

DISCUSSION

A primary goal of evolutionary biology is to determine the influences of positive and purifying selection, as well as neutral forces in shaping genetic variation. Natural populations spanning wide climatic gradients offer an ideal opportunity to investigate these patterns. We sequenced over 500 *P. trichocarpa* individuals from across much of the species range and identified over 17 million SNPs (Table 1, Fig. 2). These polymorphisms revealed significant spatial/geographic structure, even at fine scales. As previously suggested based on small-scale sequencing and genotyping[22], such patterns

appear to have resulted from a combination of restricted gene flow and complex demographic history.

Geographically structured, adaptive phenotypic variation is common among forest trees[14,24,43]. Climate is a fundamental driver of such variation[14,24,25], and we identified quantitative trait differentiation and climate-related variation within our sample consistent with this pattern. However, the molecular and evolutionary processes underlying such adaptation often remain unknown. While genome-wide polymorphism patterns suggest strong purifying selection throughout genic space, we also identified regions of the genome with unusually long haplotypes, among population differentiation, and climatic gradients consistent with recent positive or divergent selection. Genes within these regions contain a variety of annotations plausibly related to local biotic and abiotic conditions, including photoperiod-responsive and dormancy-related loci, insect and pathogen defense, abiotic stress tolerance, and phenylpropanoid metabolism. Such genes provide excellent targets for natural selection and for functional studies aimed at elucidating the drivers of local adaptation in black cottonwood and other species.

These largely non-overlapping regions also provide insight into the variety of selection pressures and modes of selection acting within and among populations. For instance, classic, recent selective sweeps (iHS) are overrepresented among genes with annotations associated with heavy metal homeostasis and symbiosis. On the other hand, if climate-driven selection primarily acts upon standing variation rather than new mutations, subtle allele frequency shifts among populations for many loci of small effects may be expected rather than hard selective sweeps. This is consistent with relatively little overlap among outlier regions identified with bayenvPC2 and iHS. Adaptation, therefore, likely

294 occurs through different process for different mutations, perhaps dependent on mutation
295 age, trait heritability and penetrance, and number of loci involved as has been suggested
296 to occur in human populations[44].

297 Remarkably, the selection outlier loci were also enriched for polymorphisms
298 associated with adaptive traits like bud flush, bud set, and height. While factors such as
299 stratification and linkage may produce erroneous associations[45], mapping traits to
300 computationally identified selection regions lends greater support to their functional
301 significance. Similar patterns have been observed in the model annual plant Arabidopsis,
302 where genomic regions showing signatures of selection are structured by climate
303 variation[9,12] and co-located with adaptive trait associations[9]. Similar examples have been
304 identified in domesticated crops[5,11]. However, to our knowledge this is the first report of
305 such concordance in a widespread, ecologically important undomesticated plant species.

306 We recognize that complex peaks of association may also be partially responsible
307 for the overlap between selection scans and GWAS and differences in GWAS signal
308 among gardens. LD combined with spurious patterns of random mutation or neutral
309 stratification may produce synthetic associations[45] and/or composite phenotypes driven
310 by multiple causal loci[46]. However, there is no reason to expect this correlative effect at
311 high frequency on a genome-wide scale. Therefore, our findings suggest that the outliers
312 contain variation relevant to adaptation based on their statistically stronger than expected
313 adaptive trait association signal.

314 The power of combining selection scans and association analyses is well
315 illustrated by insights gained from our study into winter dormancy control in natural
316 settings. Building upon previous functional studies under highly controlled

317 environments[40–42,47], our results support a model of vegetative bud set and spring bud
318 flush timing that centers on regulation of expression and symplastic mobility of the *FT1*
319 and *FT2* proteins. *FT1* is known to be transiently induced by chilling during winter and
320 promotes the floral transition[40]. However, associations of *FT1* with vegetative bud flush
321 suggest an additional function. Prolonged chilling releases endodormancy, the timing of
322 which is correlated with bud flush through subsequent accumulation of warm-
323 temperature units[24]. Moreover, the timing of the reopening of callose-plugged symplastic
324 paths, endodormancy release, and *FT1* upregulation are correlated[42]. Based on our
325 association results, we hypothesize that *FT1* is also involved in regulating endodormancy
326 release, and hence subsequent bud flush timing.

327 Reported studies of *Populus CEN1*, a flowering repressor and homolog of the FT
328 antagonist *TFL1*, also provide support for this model[48]. Its winter expression is low when
329 *FT1* expression is high, but *CEN1* is highly and transiently upregulated shortly before
330 bud flush. However, constitutive overexpression of *CEN1* delays endodormancy release
331 and bud flush[48]. In Arabidopsis, the balance between *FT* and *TFL1* appears to be central
332 to the transition to flowering versus maintenance of indeterminate meristems[49]. Thus,
333 *CEN1* might counterbalance *FT1* promotion of endodormancy release. In this model, the
334 relative timing of *FT1* regulation could influence phenotypic variation observed in bud
335 flush timing.

336 Patterns of adaptive variation are not independent of genomic history, and large-
337 scale events such as whole-genome duplications can alter the evolutionary trajectories of
338 certain loci. The deficiency of Salicoid duplicates among iHS outliers indicates that
339 recent hard selective sweeps are less likely for genes retained from genome duplication,

possibly because of fitness costs associated with altered function and/or stoichiometry of paralogs with large numbers of protein-protein interactions[50,51]. Furthermore, selective sweeps tend to affect both paralogs of a duplicated pair when they do occur, providing further support for the role of dosage constraints in duplicate gene evolution.

This is not to suggest that dosage constraints are the sole or even the primary drivers of the retention and evolution of duplicate genes. Abundant evidence supports subfunctionalization and neofunctionalization of Salicoid duplicates[31]. The case of the *FT* paralogs is again illustrative. *FT1* and *FT2* are Salicoid duplicates with divergent functions affecting distinct aspects of phenology, and displaying diametrically opposed expression patterns in *Populus*[40]. While *FT1* is primarily expressed during winter in dormant buds, *FT2* is mainly expressed during the growing season, maintaining vegetative growth[40]. Short days during fall lead to *FT2* suppression, in part through phytochrome influence on the transcription factor *PFT1*[40,41]. In support of this model, we found bud set associations with *FT2* and a *PFT1* paralog, and bud flush associations for *FT1*. This remarkable divergence in function demonstrates the adaptive potential of Salicoid duplicate pairs, consistent with classic models of duplicate gene evolution[52,53].

Intriguingly, a Salicoid duplicate pair that occurred in the CSRs are 1,3-β-glucan [callose] synthase homologs (Potri.002G058700 & Potri.005G203500). Arabidopsis callose synthases, when expressed in the phloem, deposit callose in the plasmodesmata, altering sugar and signaling molecule transport[54,55]. Returning to the phenological model outlined above, Rinne et al.[42] hypothesized the formation and degradation of callose plugs to be a control point for dormancy onset and release, possibly blocking

translocation of *FT1/FT2*. These duplicates may also have divergent functions and expression patterns, similar to those observed for the *FT* paralogs.

Our findings have important implications for understanding mechanisms of adaptation of ecologically dominant plants with widespread distributions. While forestry trials have for over 200 years indicated substantial local adaptation of dominant trees[56], ours is the first to explore the genomic legacy of this selection across the entire genome and highlight both the wide range of selection pressures as well as the climatic influence on phenological systems. These findings also have important implications for the management of natural populations in the face of environmental change. Traditionally seed transfer zone guidelines have required large numbers of plantations to accurately estimate transfer parameters[57]. Computationally identifying adaptive variants through selection scans and genome-wide phenotypic prediction could provide information in the absence of extensive plantation trials, maximizing genetic diversity while matching germplasm to current and future environmental pressures. Management and modification of such genetic diversity will undoubtedly impact dependent biotic communities and ecosystem functioning, which are known to be influenced by tree genetic variation[17].

The 17.9 million SNPs we identified represent naturally segregating variants found in wild populations, which can be utilized for multiple objectives. Forest tree improvement has traditionally relied upon natural variation in breeding programs through targeted crossing based on superior phenotypes[20]. The availability of whole-genome sequences can enable alternative breeding approaches, including genome-wide phenotypic prediction[58] and breeding with rare defective alleles, which relies on rare, recessive mutations of large effect that are commonly heterozygous and therefore masked

from many approaches[59]. Most SNPs found here are intergenic and uncommon, but many have predicted major effects in genic regions. Several SNPs of the latter type are in the candidate selection regions, including altered start and stop codons and alternative splice variants, which could represent an immediate set of tractable targets for breeding programs constrained by long generation times. Several occur at high frequency in the isolated southern or northern populations, demonstrating that sampling populations throughout the range, including marginal populations, will yield many more variants of potential utility.

Online Methods – see separate document

URL

http://www.phytozome.net/poplar.php

ACKNOWLDGEMENTS

This work was supported by funding from the BioEnergy Science Center, a U.S. DOE Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. We thank the members of BESC for their varied contributions to this work, and especially those involved in the collection, propagation, and maintenance of the common gardens, including Glenn Howe, Andrew Groover, Reinhard Stettler, Jon Johnson, and the staffs at Mt. Jefferson Farms and Greenwood Resources. We thank the WVU High Performance Computing facility, in particular Nathan Gregg and Mike Carlise. A.M.B acknowledges support from the Virginia Agricultural Experiment Station and the Program McIntire Stennis of the National Institute of Food and Agriculture, U.S. Department of Agriculture.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests

Author Contributions:

G.A.T., S.P.D., G.T.S., & L.M.E. conceived and designed the study. All authors performed measurements. L.G., J.M., & W.S. performed sequencing. L.M.E., S.P.D., G.T.S., E. R.-M., J.M., P.R., W.M., & W.S. performed analyses. L.M.E., S.P.D. and A.M.B. drafted the manuscript. All authors read, revised, and approved the manuscript.

FIGURE LEGENDS:

Figure 1. Geographic locations and genetic structure of the 544 *P. trichocarpa* individuals sequenced. **a.** Map of collection locations of the 544 *P. trichocarpa* genotypes sampled in this study from along the Northwest coast of North America, with the species range shaded in tan, and PCA of all 544 individuals color-coded by general geographic regions. Yellow diamonds represent plantation locations. **b.** PCA of the central WA/BC group of individuals (outlined by box in part (a)) color-coded by collection river. The percent of the variance explained by the first two PC axes for both the regional analysis and the WA/BC group is shown.

Figure 2. Phenotypic evidence of climate-driven selection in *P. trichocarpa*. **a.** Patterns of quantitative trait differentiation (Q_{ST}) are stronger than genome-wide differentiation (F_{ST}) among sampled geographic locations. Shaded area represents the 95% confidence interval (CI) of F_{ST}, while points and bars represent the point and 95% CI of Q_{ST}. **b-d.** Genotypic estimates of best linear unbiased predictors for adaptive traits growing in multiple plantation environments show strong correlations with the first principal component of 20 climate variables measured at the collection location. Negative PC1 values are associated with warmer conditions, while more positive bud flush and bud set

BLUPs indicate more earlier flush or set, respectively. Correlation coefficient and p-value are shown above each.

Figure 3. Unique and shared genomic regions among five selection scans. **a.** A Venn diagram of the number of regions throughout the genome in the top 1% for each selection scan. **b.** Overrepresentation p-value for panther annotation categories in selection outliers. Only the 10 most strongly overrepresented categories for each selection scan are shown.

Figure 4. The selection outliers have a stronger association signal with adaptive traits than expected by chance. **a-c.** The genome-wide distribution of association signal in 1-kb windows through the genome (blue; left axis) and the association within the selection outliers (green; right axis; red line indicates mean) for three traits in different gardens.

Figure 5. A region of chromosome 10 that displays an abundance of bud flush association and strong evidence of selection from multiple different selection scans. Dashed lines represent the 1% cutoff mark for selection scans.

Figure 6. A region of chromosome 8 that displays multiple strong bud flush associations, in addition to evidence of positive selection. Dashed lines represent the 1% cutoff mark for selection scans.

Table 1. Per-site nucleotide diversity, π, estimated across the genome for all annotated features of the *P. trichocarpa* v3 genome, and the number of variants annotated in each class using SnpEff[60].

Feature	π (median and central 95% range)
Overall	0.0041 (0.0004-0.01226)
Intergenic	0.0064 (0.0012-0.0125)
Genic[a]	0.003 (0.0006-0.0106)
5'UTR	0.0028 (0.0001-0.0114)
3'UTR	0.0033 (0.0001-0.0123)
Intron	0.0034 (0.0005-0.0114)
Coding Sequence	0.002 (0.0002-0.0111)
Nonsynonymous	0.0018 (0-0.0122)
Synonymous	0.0054 (0-0.0348)
π_{Nonsyn}/π_{Synon}	0.3179 (0-14.5447)

Annotation	Number of variants[b]
Intergenic	14,520,224
Intron	1,962,848
Non-synonymous coding	612,655
Non-synonymous start	253
Start lost	1631
Stop gained	18,702
Stop lost	2175

Splice site acceptor	3748
Splice site donor	4449
Synonymous coding	386,103
Synonymous stop	959
3’ UTR	389,771
5’ UTR	169,083

458

459 [a] Predicted transcript from 5’ to 3’UTR

460 [b] Total is greater than total observed number of variants because some SNPs have

461 multiple annotations for alternative transcripts

462

463

Table 2. Tests of over- and underrepresentation of retained Salicoid duplicate genes and pairs among the selection outliers. Shown are the number of genes in each category and the associated p-value. 39,514 genes are found on the 19 chromosomes, with 7609 pairs from 15,797 genes.

Selection Scan	Duplicate Genes in Outlier Regions	Fisher's Exact Test (p-value)[a]	Duplicate Pairs in Outlier Regions	Fisher's Exact Test (p-value) [a]
CSR	178	NS (0.623)	2	NS (0.263)
F_{ST}	674	Over ($2.8x10^{-9}$)	27	Over (0.002)
SPA	741	Over (0.004)	24	NS (0.065)
iHS	348	Under ($3.0x10^{-12}$)	8	Over (0.039)
BFPC1	100	NS (0.661)	1	NS (0.263)
BFPC2	134	NS (0.156)	0	NS (1)

[a] NS, not significant; Over or Under, genes or pairs were significantly overrepresented or underrepresented within outlier regions, respectively, compared to genome-wide expectation.

Methods

Sequencing, assembly, and variant calling

We obtained plant materials from 1100 black cottonwood (*Populus trichocarpa* Torr & Gray) from wild populations in California, Oregon, Washington, and British Columbia, as previously described[22]. We resequenced a set of 649 genotypes to a minimum expected depth of 15x using the Illumina Genome Analyzer, HiSeq 2000, and HiSeq 2500. Sequences were down-sampled for those individuals sequenced at greater depths to ensure even coverage throughout the population (Supplementary Fig. 1a). Short reads were then aligned to the *P. trichocarpa* version 3 genome using BWA 0.5.9-r16 with default parameters[61]. We corrected mate pair metadata and marked duplicate molecules using the FixMateInformation and MarkDuplicates methods in the Picard package (http://picard.sourceforge.net). Next, we called SNPs and small indels for the merged dataset using SAMtools mpileup (-E –C 50 –DS –m 2 –F 0.000911 –d 50000) and bcftools (-bcgv –p 0.999089)[62].

Genotype validation

We compared the samtools mpileup genotype calls for 649 individuals to 22,438 SNPs assayed on the *Populus* Illumina Infinium platform, which was designed based on assembly version 2.0[22,63]. These were high-quality SNPs that we could confidently place on the v3 reference genome. The 649 individuals had, on average, a 97.9% match rate. SNPs with a minor allele frequency (MAF) ≥ 0.05 had a match rate of 98.1%, while those with MAF ≤ 0.01 (n=159 SNPs) had a match rate of 78.2%, similar to other published studies[4,64,65]. Stringent filtering had minimal impact on match rate, though it reduced substantially the number of known SNPs passing the filtering thresholds. For example, requiring an individual minimum depth of 3, minimum mapping quality of 30, minor allele count of 15, and minimum quality score of 30 increased the false negative rate by 3.9%, but only increased the match rate by 0.3%. Therefore, no additional filtering after samtools mpileup variant calling was performed.

Nisqually-1 was the original individual sequenced by Tuskan et al.[29] using Sanger technology, and it was also resequenced during this study using the Illumina platform. 716,691 heterozygous polymorphisms found in the v3.0 reference genome assembly (http://www.phytozome.net/poplar.php) had at least three Sanger reads of each allele, and therefore had strong evidence of being heterozygous in the Sanger assembly. In the current study, we correctly identified 557,738 of these (77.82%), including 3,205 of 3,220 singleton variants in Nisqually-1 in the Illumina data, suggesting a 22.18 % false negative rate. Conversely, of 1,115,963 heterozygous positions identified in Nisqually-1 in the current Illumina genotyping, 972,254 had at least one Sanger read supporting each allele, suggesting a 12.86 % false positive rate. All of these comparisons were done with no filtering of the samtools mpileup genotype calls. It is important to note that errors occur in both the Sanger and Illumina methods, so these are likely to be overestimates of the true error rates in the resequencing SNP data.

The Accessible Genome

Next, we identified the *Populus trichocarpa* "accessible genome" as those positions that had sufficient read depth across enough individuals to enable genotypes to be accurately determined (similar to the approach used in the 1000 Genomes Project[1]).

We estimated the median and interquartile range of depth for each position in the genome, for all sequenced individuals, using samtools mpileup. With our target of 15X coverage, "accessible" positions were those with median depth between 5 and 45 (inclusive) and with an interquartile range less than or equal to 15 (Supplementary Fig. 1a,b). Of the 394,507,732 positions that were sequenced across all individuals, 345,217,484 met these criteria (~87.51%), 17,902,170 of which were single nucleotide polymorphisms (SNPs) (15,454,190 biallelic). We observed a slight deficiency of heterozygotes at lower depth positions; however, these positions cumulatively comprise only between 0.7 and 2.5% of positions at an uncorrected HWE p-value threshold of 0.001 (Supplementary Fig. 1c). Furthermore, these cutoffs did not bias the outcomes of selection scans throughout the genome, as putative selection outliers (see below) had a very similar distribution of depth as the rest of the genome (Supplementary Table 14) and there was no relationship with association p-value (see below; all Pearson $|r| < 0.005$, Supplementary Fig. 1d).

Relatedness, Hybridization, and Population Structure

We next identified individuals that showed evidence of admixture with other species of *Populus* because hybridization is common within the genus[66]. We used 7 additional individuals sequenced to at least 32X depth as above: 3 *P. deltoides*, 1 *P. fremontii*, 1 *P. angustifolia*, 1 *P. nigra*, and 1 *P. tremuloides*. These were aligned to the *P. trichocarpa* v3.0 reference genome using Bowtie2 in local alignment mode and default parameters[67], and variants were called using the samtools mpileup function for each species separately. We then used smartpca[68] to identify sampled individuals in this study that were genetically similar to these alternative species. This method identified 3 individuals that appear intermediate between the *P. trichocarpa* cluster and an alternate species (Supplementary Fig. 14).

We performed similar analyses using overlapping genomic regions from 32 *P. balsamifera* transcriptomes (provided courtesy of Dr. Matt Olson, Texas Tech University; Supplementary Fig. 15), and, separately, the Illumina Infinium array data, which contained additional individuals of alternative species[63]. These identified an additional three genetically intermediate individuals. These 6 potentially admixed individuals were removed from subsequent analyses.

We next identified and removed individuals more related than first cousins using the program GCTA[69]. Because this, like most other relatedness estimates, relies on allele frequency estimates within populations, it was necessary to first identify genetic clusters. We iteratively identified genetic clusters using PCA[68], each representing a putative genetic group. We removed related individuals within each from further analyses, leaving a total of 544 individuals, which were used for all subsequent analyses.

To assess population structure, we used PCA analyses with these unrelated 544 individuals. This identified roughly 4 major groupings (Figure 1a). We then performed PCA analysis using only those individuals from the Washington/British Columbia group to investigate finer-scale structure (Fig. 1b). PCA was performed using all 17.9 million SNPs.

Phenotypic Evidence of Selection

We investigated phenotypic evidence of selection using two methods. First, we compared neutral genetic differentiation among collection rivers/subpopulations (F_{ST}, see

below for details of estimation) to differentiation among rivers for second-year height and fall and spring phenology using data collected from three replicated plantations (Q_{ST}). Briefly, over 1000 *P. trichocarpa* genotypes were planted in 2009 in three replicated common gardens (Clatskanie and Corvallis, OR, and Placerville, CA) in a randomized block design with three replicates of each genotype. In 2010, we measured spring bud flush, fall bud set, and total height in each garden. We removed within-garden micro-site variation using thin-plate spline regression (*fields* R package), then estimated among river, among genotypes within rivers, and residual variance components (σ^2_R, σ^2_G, and σ^2_ε, respectively) using mixed-model regression (*lmer* function of the *lme4* R package). Q_{ST} was estimated at the river level as $\sigma^2_R/(\sigma^2_R + 2*\sigma^2_G)$[32]. A 95% confidence interval of Q_{ST} was estimated by resampling rivers, with replacement, 1,000 times and estimating Q_{ST} for each bootstrapped dataset. We directly compared the 95% CIs for Q_{ST} and F_{ST}. We note that in using clonal replicates σ^2_G includes additive and non-additive genetic effects, rather than the additive genetic variance alone; however, simulations have shown that this approach lowers Q_{ST} estimates, and is therefore a conservative test of $Q_{ST} > F_{ST}$[70].

Second, we tested for correlations between these adaptive traits and the climate of the source location. We tested correlations with mean annual temperature, mean annual precipitation, and the first two principal components (cumulatively > 85% of variance explained) of 20 climate variables obtained using ClimateWNA[71]. We used the genotypic best linear unbiased predictors obtained from mixed model analysis (*lmer* function of the *lme4* R package) as the phenotypic traits. Climate variables were averaged within collection locations prior to correlation analysis.

Genetic Variation and Signatures of Recent Positive Selection Throughout the Genome

We assessed species-wide nucleotide diversity (π)[72] using the MLE estimate of allele frequency from the samtools mpileup output[62] in all annotated regions (coding sequence, introns, 5' and 3' UTRs) of the v3 genome greater than 150 bp long and with at least 95% accessibility.

We performed five genome-wide scans of recent positive selection, using four conceptually different approaches. First, we estimated genetic differentiation[72] among collection rivers as F_{ST} in 1-kb windows throughout the genome (again, requiring at least 95% accessibility and using the accessible positions in a window as the window's full length). We restricted this analysis to rivers/subpopulations with at least eight individuals, and randomly chose 20 individuals from those that contained > 20 individuals (14 rivers total: Homathko, Skwawka, Lillooet, Squamish, Salmon, Fraser, Columbia, Nisqually, Nooksack, Puyallup, Skagit, Skykomish, Tahoe, Willamette). We estimated nucleotide diversity across all individuals (π_T) and weighted within-river nucleotide diversity (π_S), accounting for sequencing error[73]. We calculated F_{ST} as difference between total and weighted within-river diversity, divided by the total diversity (π_{T-S} / π_T)[72]. We took the top 1% of the empirical distribution of F_{ST} as genomic regions representing unusually strong allele frequency differences among rivers and candidates of divergent selection.

The second selection scan quantified the steepness of allele frequency clines across two climate variables, using the program SPA[33]. SPA uses a logistic regression-based approach to model allele frequency clines, without *a priori* population assignment and represents a fundamentally different approach than the F_{ST} scan described above. We

used mean annual temperature and mean annual precipitation of the source location for each sample, obtained using ClimateWNA[71], because these variables are significantly correlated with growth and phenological traits. We averaged SPA in non-overlapping 1-kb bins throughout the genome, requiring at least 5 SNPs in each window. We identified the top 1% of these windows as regions of the genome with unusually steep allele frequency clines across mean annual temperature and precipitation.

Third, we identified regions of the genome with recent, unusually rapid increases in allele frequency across the range. Strong, recent selective sweeps will result in long haplotypes associated with the selected allele[8,74]. First, we phased the 544 diploid individuals using SHAPEIT2[75]. Because we have no reference haplotype panels to test the accuracy of computationally-determined haplotypes, we determined the optimal method by estimating the accuracy of imputed masked loci[76]. We used 10 Mb of chromosome 2 (5-15Mb), using only variants with MAF>0.1 (307,123 sites). We randomly masked out 5% of the center 260,000 positions for each individual (avoiding the ends), treating them as missing for phasing. To determine the optimal number of hidden Markov states (K) and the window size (W) used in SHAPEIT2, we phased the data using combinations of parameters from K=50-600 and W=0.1-2Mb (Supplementary Fig. 14), using the default N_e=15K, and run with 4 threads. The genetic position was determined through linear interpolation using a genetic map derived from a *P. trichocarpa* x *P. deltoides* pseudo-backcross pedigree and 3,559 Infinium SNP markers[22]. Genetic position and recombination rate were estimated using local linear regression with the *loess* function in R. For comparison, we also phased the same data using the default settings of BEAGLE[77]. We then determined the squared correlation coefficient (R^2) between the known allele dosages (0, 1, or 2) and the imputed genotypes for masked positions in each individual. The average R^2 is shown in Supplementary Fig. 16, and peaks at approximately K=350, W=0.1 Mb. We varied N_e from 10,000 – 20,000, and found that Ne=15,000 gave the highest correlation between known and imputed allele dosage for masked missing data. Using the same 10Mb region of chromosome 2, we tested whether the 0.1 MAF cutoff affected accuracy, and found that with no MAF cutoff accuracy was actually increased. We therefore phased all chromosomes using SHAPEIT2 with K=350 states, W=0.1 Mb window size, and Ne=15,000 effective population size, using all non-singleton and -private doubleton sites, parallelized using 24 threads.

We then estimated the integrated haplotype score (iHS[8]) for SNPs. Because the program is computationally intensive, we thinned the dataset to SNPs separated by at least 100bp and with a MAF of at least 0.05, resulting in 1,898,506 SNPs throughout the genome. In calculating iHS, we used the genetic distance as described above. iHS was standardized within allele frequency bins[8], and |iHS| averaged within non-overlapping 1-kb windows, again requiring at least 5 SNPs in a window. We took the top 1% of these bins as genomic regions that have experienced an unusually rapid allele frequency change, resulting in extended haplotype homozygosity, and potential targets of positive selection.

Finally, we used bayenv2.0[34] to identify regions of the genome with unusually strong allele frequency clines along climatic gradients while controlling for background neutral population structure. We performed this analysis with 13 of the populations used in the F_{ST} analysis described above. We excluded the Tahoe population because it was so divergent that the neutral model of bayenv2.0 had difficulty accounting for the covariance in allele frequencies among populations (data not shown). We used the first

two principle components (PCs) of the climate data from source locations, averaged within populations, which cumulatively explained >85% of the variance in the correlation matrix. Loadings showed that the first PC was strongly related to all climateWNA variables, while the second PC was more strongly related to precipitation, heat-moisture indices, and frost free period metrics (Supplementary Fig. 17). To estimate the covariance matrix of allele frequency among populations, we used 19,420 genome-wide SNPs that were separated by at least 20Kbp and with MAF > 0.01 across the 13 populations using bayenv2.0 with 100,000 steps through the chain, performed three times independently. The three runs were very similar (all Mantel R > 0.985, $p<0.001$), and the difference in covariances among runs were always less than 3% of the smallest estimated covariance, indicating convergence[78]. We assessed the strength of the correlation of allele frequency and the climate variables, as estimated by the Bayes factor (BF) and Spearman correlation, for 9,519,343 SNPS (MAF > 0.01 across the 13 populations). We tested, for 20,000 randomly-chosen SNPs, the effect of chain length on the Bayes factors. Correlations of the individual SNPs among the different chain lengths and independent runs for each chain length indicated that 10 chains of 50,000 steps were sufficient to ensure repeatability and accuracy (Supplementary Fig. 18), while tractable for millions of SNPs. For the final analysis of all >9.5million SNPs, we calculated the Bayes factor and Spearman correlation using 50,000 steps in each of 10 independent runs. We averaged the $\log_{10}$(BF) and the posterior Spearman correlation estimate for each SNP, normalized these values within MAF bins (0.05 bin size), and averaged these within 1-kb windows throughout the genome, requiring at least 5 SNPs per 1-kb window.

To identify regions of the genome with unusually strong allele frequency-climate correlations, we selected the windows in the top 1% of Spearman climate-allele frequency correlations and top 1% of Bayes Factors as those with unusually strong climate related allele frequency clines. This process was done separately for the first and second PCs, resulting in two separate selection scans.

Candidate Selection Regions (CSRs) and Annotation Analysis

The selection scans represent five different approaches to identifying unusually strong patterns throughout the genome that are consistent with recent positive or divergent selection. Merging nearby windows (≤5Kb), we found 397 regions that were in the top 1% of at least two of the five scans (the candidate selection regions, or "CSRs"), spanning or adjacent to 452 different genes. We identified the genes spanning or nearest to the CSRs and selection outlier regions. We used Fisher Exact Tests to determine if GO, PANTHER, and PFAM annotations were overrepresented in the genes associated with the CSRs and outlier regions.

We also tested whether these genes were overrepresented among lists from known gene families and pathways, and known to be responsive to drought and dormancy cycling. Families of transcription factors were identified using the Plant Transcription Factor Database v3.0 (http://planttfdb.cbi.pku.edu.cn/index.php?sp=Pth[79]). Genes in additional pathways and families are listed in Supplementary Table 11. When necessary, we used the best reciprocal BLAST hit between the v1 and v3 genome assemblies to locate the gene models identified by previous studies for each set of published genes.

Genome Duplication and Network Connectedness

First, we examined the genes spanning or nearest to the CSRs and the windows of the top 1% of each selection scan in the context of the Salicoid whole-genome duplication using the 7,936 duplicate pairs identified by Rodgers-Melnick et al.[31]. We used Fisher Exact Tests (FET) to test whether these selection scan lists were under- or over-represented among the duplicate pairs. To determine if there were more duplicate pairs in which both genes of the pair were associated with the selection outliers than expected by chance, we used a random resampling procedure. For each selection scan, we resampled without replacement the same number of genes observed in that scan that were also retained duplicates from the total number of retained duplicates (15,812) 10,000 times and recoded how many complete pairs were resampled each time, meaning how many times both genes of a pair were randomly sampled. We tested whether genes associated with selection outliers had more protein-protein interactions (PPI) than expected. We used the number of connections in protein-protein interaction networks with 65 % confidence determined by the ENTS random forest prediction program[30]. We tested whether PPIs of the genes in each scan were different from the genome-wide average using Wilcoxon two-sample tests. These analyses examined patterns of genes associated with the CSRs and the selection outlier regions.

We also examined patterns at the whole-gene level, by calculating π_S, π_T, and the ratio of nonsynonymous/synonymous polymorphism ($\pi_{Nonsynonymous}/\pi_{Synonymous}$) for 39,514 genes on the 19 chromosomes using the same methods described above. We then calculated the correlation of each statistic between the 7,936 Salicoid duplicate pairs of genes. To determine if the observed correlation was greater than expected by chance, we randomly chose 7,936 pairs of genes from all genes 10,000 times, as a null distribution of correlation between pairs of randomly chosen genes.

We also tested whether the mean observed selection statistic differed between Salicoid duplicates and non-duplicate gene using Wilcoxon two sample tests. To test whether the connectedness of genes may influence patterns of selection, we examined correlations between PPI and the observed statistics. We assessed significance using 10,000 permutations of connectedness across the test statistic as above. We $\log_{10}$-transformed the data as necessary.

Signal of Association Throughout the Entire Genome and Within the CSRs

To determine if loci within the identified regions may have functional significance, we tested for statistical associations with second-year height and fall and spring bud phenology using data collected from three replicated plantations. We estimated genotypic best linear unbiased predictors using mixed-model regression (*lmer* function of the *lme4* R package, see Phenotypic Selection section above) as the phenotypes for GWAS. We used the same set of resequenced, unrelated individuals used described above, excluding the highly differentiated Tahoe, Willamette Valley, and far northern British Columbia samples because strong stratification can lead to spurious associations[80], leaving 498 individuals. We only tested phenotypic association with SNPs having a MAF≥0.05, leaving 5,939,334 SNPs. The analysis was performed for single traits in each plantation using emmax[36], using the IBS kinship matrix to account for background genetic effects. To account for population structure, for each trait we included as covariates the principal components axes that were significant predictors of the trait, chosen using stepwise regression (*step* function in the R package). We used the

gemma multi-trait association model[37] to test for SNP association with each trait across all three plantations simultaneously, and in a 9-trait model as well (3 traits x 3 plantations). We used the mixed-model framework incorporating kinship and principal component axes that were significant (nominal alpha=0.05) in a multivariate multiple linear regression.

We estimated alpha values for association p-values by permutation[81]. We permuted individual alleles among individuals, randomly generating genotypes while mirroring exactly the true MAF distribution. We then tested for association of these random genotypes with the observed phenotype data using the actual kinship matrix and principal components as above, thereby testing only the effect of randomly assigned genotypes while the structure of population stratification, relatedness, and the phenotypes was held constant. For univariate analyses performed in emmax we performed 10^8 permutations. For gemma multi-trait analyses, we used $>10^8$ permutations for bud set and height and 8-33x10^6 permutations for bud flush and the 9-trait model, which were computationally more intensive. For each trait, we then estimated the cutoffs at various alpha levels (Supplementary Table 15).

To determine if the observed associations within the selection outliers was greater than expected by chance, we used the $-\log_{10}$(p-value) as the association signal within each selection outlier, and used the average of these values for each trait. We then randomly sampled the same number of 1-kb bins from throughout the genome 20,000 times. The number of random samples with a mean equal to or greater than the observed for each trait represents the probability of finding a median association signal in the selection outliers by chance alone. We also calculated the empirical p-value for each CSR using the distribution of average association p-values within 1-kb windows throughout the genome. This was done while controlling for the distribution of gene density within the surrounding 100 kb of the selection scans (Supplementary Figure 11g). We also repeated this with a 50-kb window and without controlling for gene density, and found the same patterns (data not shown).

References

1. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

2. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43,** 956–963 (2011).

3. Axelsson, E. *et al.* The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495,** 360–364 (2013).

4. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44,** 808–811 (2012).

5. Huang, X. *et al.* Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44,** 32–39 (2012).

6. Zhao, S. *et al.* Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat. Genet.* **45,** 67–71 (2013).

7. Miller, W. *et al.* Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2382–2390 (2012).

8. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).

9. Fournier-Level, a *et al.* A map of local adaptation in *Arabidopsis thaliana*. *Science* **334,** 86–89 (2011).

10. Tishkoff, S. a *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39,** 31–40 (2007).

11. Jia, G. *et al.* A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45,** 957–961 (2013).

12. Hancock, A. M. *et al.* Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334,** 83–86 (2011).

13. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152,** 703–713 (2013).

14. Savolainen, O., Pyhäjärvi, T. & Knürr, T. Gene Flow and Local Adaptation in Trees. *Annu. Rev. Ecol. Evol. Syst.* **38,** 595–619 (2007).

15. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320,** 1444–1449 (2008).

16. Ellison, A. M. *et al.* Loss of foundation species : consequences for the structure and dynamics of forested ecosystems. *Front. Ecol. Environ.* **3,** 479–486 (2005).

17. Whitham, T. G. *et al.* Extending genomics to natural communities and ecosystems. *Science* **320,** 492–495 (2008).

18. Parmesan, C. Ecological and Evolutionary Responses to Recent Climate Change. *Annu. Rev. Ecol. Evol. Syst.* **37,** 637–669 (2006).

19. Ingvarsson, P. K., García, M. V., Hall, D., Luquez, V. & Jansson, S. Clinal variation in phyB2, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (*Populus tremula*). *Genetics* **172,** 1845–1853 (2006).

20. Neale, D. B. & Kremer, A. Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* **12,** 111–122 (2011).

21. Jansson, S. & Douglas, C. J. *Populus*: a model system for plant biology. *Annu. Rev. Plant Biol.* **58,** 435–458 (2007).

22. Slavov, G. T. *et al.* Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol.* **196,** 713–725 (2012).

23. Pauley, S. S. & Perry, T. O. Ecotypic variation in the photoperiodic response in *Populus*. *J. Arnold Arbor.* **35,** 167–188 (1954).

24. Howe, G. T. *et al.* From genotype to phenotype : unraveling the complexities of cold adaptation in forest trees 1. *Can. J. Bot.* **1266,** 1247–1266 (2003).

25. McKown, A. D. *et al.* Geographical and environmental gradients shape phenotypic trait variation and genetic structure in Populus trichocarpa. *New Phytol.* **201,** 1263–1276 (2014).

26. Wegrzyn, J. L. *et al.* Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytol.* **188,** 515–532 (2010).

27. Porth, I. *et al.* Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytol.* **200,** 710–726 (2013).

28. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18,** 1944–1954 (2008).

29. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).

30. Rodgers-Melnick, E., Culp, M. & DiFazio, S. P. Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. *BMC Genomics* **14,** 608 (2013).

31. Rodgers-Melnick, E. *et al.* Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Res.* **22,** 95–105 (2012).

853 32. Spitze, K. Population structure in *Daphnia obtusa*: quantitative genetic and
854 allozymic variation. *Genetics* **135,** 367–374 (1993).

855 33. Yang, W.-Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for
856 analysis of spatial structure in genetic data. *Nat. Genet.* **44,** 725–731 (2012).

857 34. Günther, T. & Coop, G. Robust identification of local adaptation from allele
858 frequencies. *Genetics* **195,** 205–220 (2013).

859 35. Sun, J., Xie, D., Zhao, H. & Zou, D. Genome-wide identification of the class III
860 aminotransferase gene family in rice and expression analysis under abiotic
861 stress. *Genes Genomics* **35,** 597–608 (2013).

862 36. Kang, H. M. *et al.* Variance component model to account for sample structure
863 in genome-wide association studies. *Nat. Genet.* **42,** 348–354 (2010).

864 37. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms
865 for genome-wide association studies. *Nat. Methods* **11,** 407–409 (2014).

866 38. Ruttink, T. *et al.* A molecular timetable for apical bud formation and dormancy
867 induction in poplar. *Plant Cell* **19,** 2370–2390 (2007).

868 39. Werner, A. K. *et al.* The ureide-degrading reactions of purine ring catabolism
869 employ three amidohydrolases and one aminohydrolase in Arabidopsis,
870 soybean, and rice. *Plant Physiol* **163,** 672–681 (2013).

871 40. Hsu, C.-Y. *et al.* FLOWERING LOCUS T duplication coordinates reproductive
872 and vegetative growth in perennial poplar. *Proc. Natl. Acad. Sci. U. S. A.* **108,**
873 10756–10761 (2011).

874 41. Iñigo, S., Alvarez, M. J., Strasser, B., Califano, A. & Cerdán, P. D. PFT1, the
875 MED25 subunit of the plant Mediator complex, promotes flowering through
876 CONSTANS dependent and independent mechanisms in Arabidopsis. *Plant J.*
877 **69,** 601–612 (2012).

878 42. Rinne, P. L. H. *et al.* Chilling of dormant buds hyperinduces FLOWERING
879 LOCUS T and recruits GA-inducible 1,3-beta-glucanases to reopen signal
880 conduits and release dormancy in *Populus*. *Plant Cell* **23,** 130–146 (2011).

881 43. Hall, D. *et al.* Adaptive population differentiation in phenology across a
882 latitudinal gradient in European aspen (*Populus tremula*, L.): a comparison of
883 neutral markers, candidate genes and phenotypic traits. *Evolution* **61,** 2849–
884 2860 (2007).

885 44. Pritchard, J. K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nat. Rev.*
886 *Genet.* **11,** 665–667 (2010).

45. Platt, A., Vilhjálmsson, B. J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186,** 1045–1052 (2010).

46. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465,** 627–631 (2010).

47. Bohlenius, H. *et al.* CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* **312,** 1040–1043 (2006).

48. Mohamed, R. *et al. Populus* CEN/TFL1 regulates first onset of flowering, axillary meristem identity and dormancy release in Populus. *Plant J.* **62,** 674–688 (2010).

49. Jaeger, K. E., Pullen, N., Lamzin, S., Morris, R. J. & Wigge, P. A. Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. *Plant Cell* **25,** 820–833 (2013).

50. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60,** 433–53 (2009).

51. Birchler, J. A. & Veitia, R. A. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* **186,** 54–62 (2010).

52. Lynch, M. & Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154,** 459–473 (2000).

53. Taylor, J. S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* **38,** 615–643 (2004).

54. Vatén, A. *et al.* Callose biosynthesis regulates symplastic trafficking during root development. *Dev. Cell* **21,** 1144–1155 (2011).

55. Xie, B., Wang, X., Zhu, M., Zhang, Z. & Hong, Z. CalS7 encodes a callose synthase responsible for callose deposition in the phloem. *Plant J.* **65,** 1–14 (2011).

56. Langlet, O. Two hundred years of genecology. *Taxon* **20,** 653–722 (1971).

57. Wang, T., O'Neill, G. a & Aitken, S. N. Integrating environmental and genetic effects to predict responses of tree populations to climate. *Ecol. Appl.* **20,** 153–163 (2010).

58. Grattapaglia, D. & Resende, M. D. V. Genomic selection in forest tree breeding. *Tree Genet. Genomes* **7,** 241–255 (2010).

59. Vanholme, B. *et al.* Breeding with rare defective alleles (BRDA): a natural *Populus nigra* HCT mutant with modified lignin as a case study. *New Phytol.* **198,** 765–776 (2013).

60. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).

61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

62. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

63. Geraldes, A. *et al.* A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Mol. Ecol. Resour.* **13,** 306–323 (2013).

64. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42,** 961–7 (2010).

65. Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44,** 812–815 (2012).

66. Eckenwalder, J. E. in *Biol. Popul. Its Implic. Manag. Conserv.* (Stettler, R. F., Bradshaw, H. D. J., Heilman, P. E. & Hinckley, T. M.) 7–32 (NRC Research Press, 1996).

67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–360 (2012).

68. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

69. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).

70. Goudet, J. & Büchi, L. The effects of dominance, regular inbreeding and sampling design on Q(ST), an estimator of population differentiation for quantitative traits. *Genetics* **172,** 1337–1347 (2006).

71. Wang, T., Hamann, A., Spittlehouse, D. L. & Murdock, T. Q. ClimateWNA—High-Resolution Spatial Climate Data for Western North America. *J. Appl. Meteorol. Climatol.* **51,** 16–29 (2012).

72. Charlesworth, B. Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* **15,** 538–543 (1998).

73. Johnson, P. L. F. & Slatkin, M. Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* **25,** 199–206 (2008).

74. Sabeti, P., Reich, D. & Higgins, J. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).

75. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).

76. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12,** 703–14 (2011).

77. Browning, B. & Browning, S. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* 210–223 (2009).

78. Pyhäjärvi, T., Hufford, M. B., Mezmouk, S. & Ross-Ibarra, J. Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* **5,** 1594–609 (2013).

79. Zhang, H. *et al.* PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.* **39,** D1114–1117 (2011).

80. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11,** 459–463 (2010).

81. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32,** 227–234 (2008).

Figure 1.

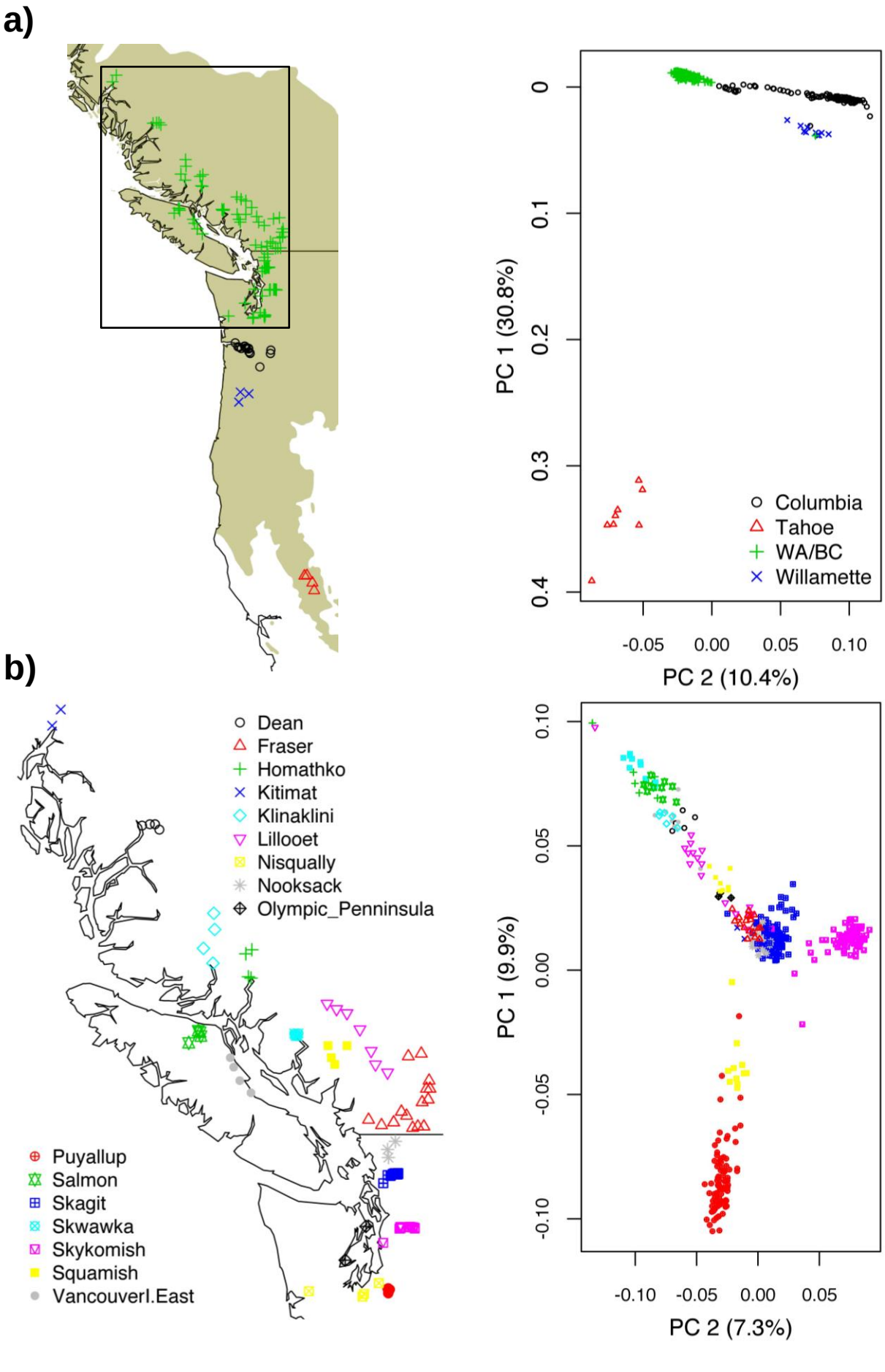

a)
PC 1 (30.8%)
PC 2 (10.4%)
Columbia
Tahoe
WA/BC
Willamette
b)
Dean
Fraser
Homathko
Kitimat
Klinaklini
Lillooet
Nisqually
Nooksack
Olympic_Penninsula
Puyallup
Salmon
Skagit
Skwawka
Skykomish
Squamish
VancouverI.East
PC 1 (9.9%)
PC 2 (7.3%)

Figure 2.

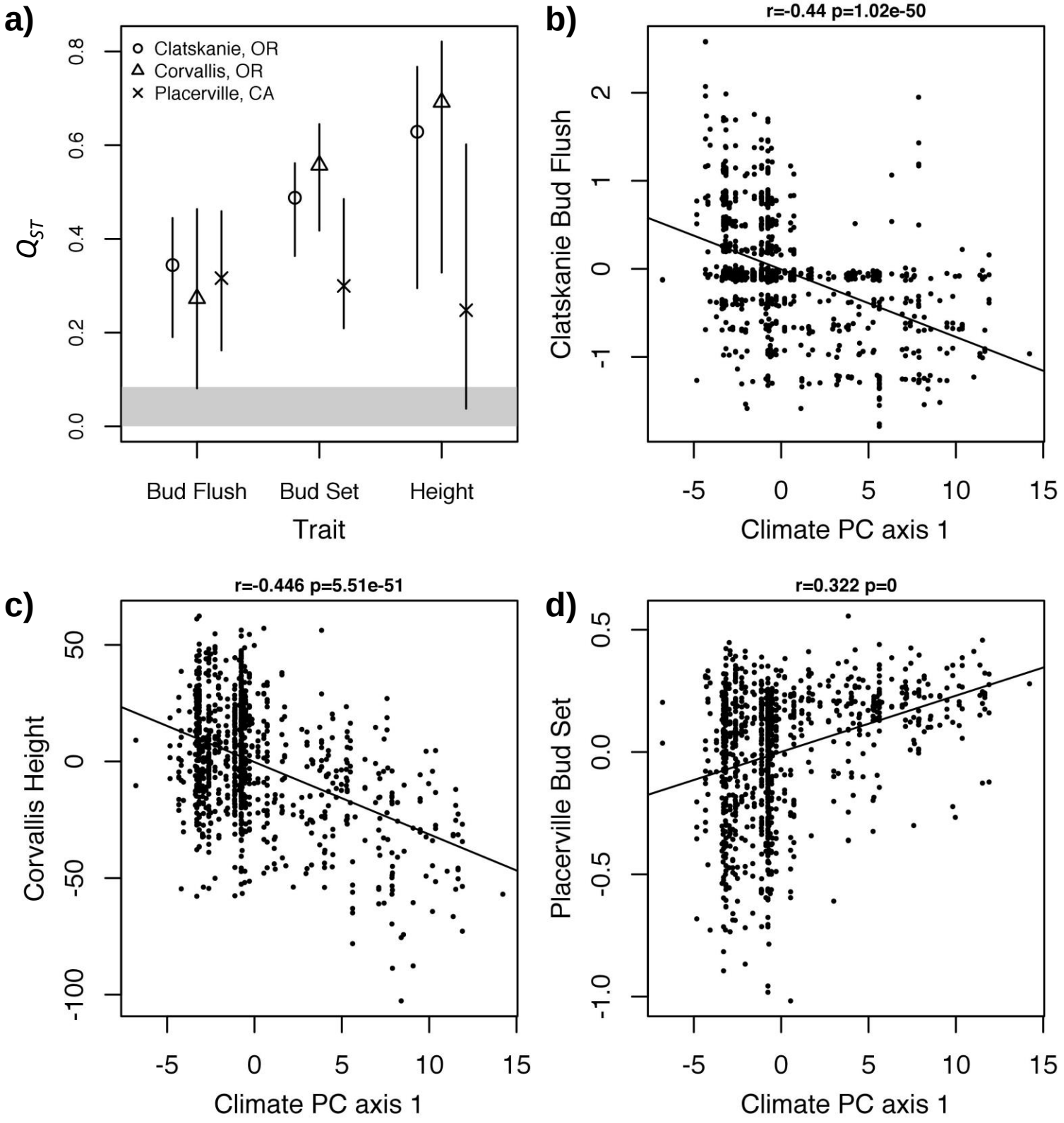

a)
Clatskanie, OR
Corvallis, OR
Placerville, CA
Q_{ST}
0.0
0.2
0.4
0.6
0.8
Bud Flush
Bud Set
Height
Trait
b)
r=-0.44 p=1.02e-50
Clatskanie Bud Flush
-1
0
1
2
-5
0
5
10
15
Climate PC axis 1
c)
r=-0.446 p=5.51e-51
Corvallis Height
-100
-50
0
50
-5
0
5
10
15
Climate PC axis 1
d)
r=0.322 p=0
Placerville Bud Set
-1.0
-0.5
0.0
0.5
-5
0
5
10
15
Climate PC axis 1

Figure 3.

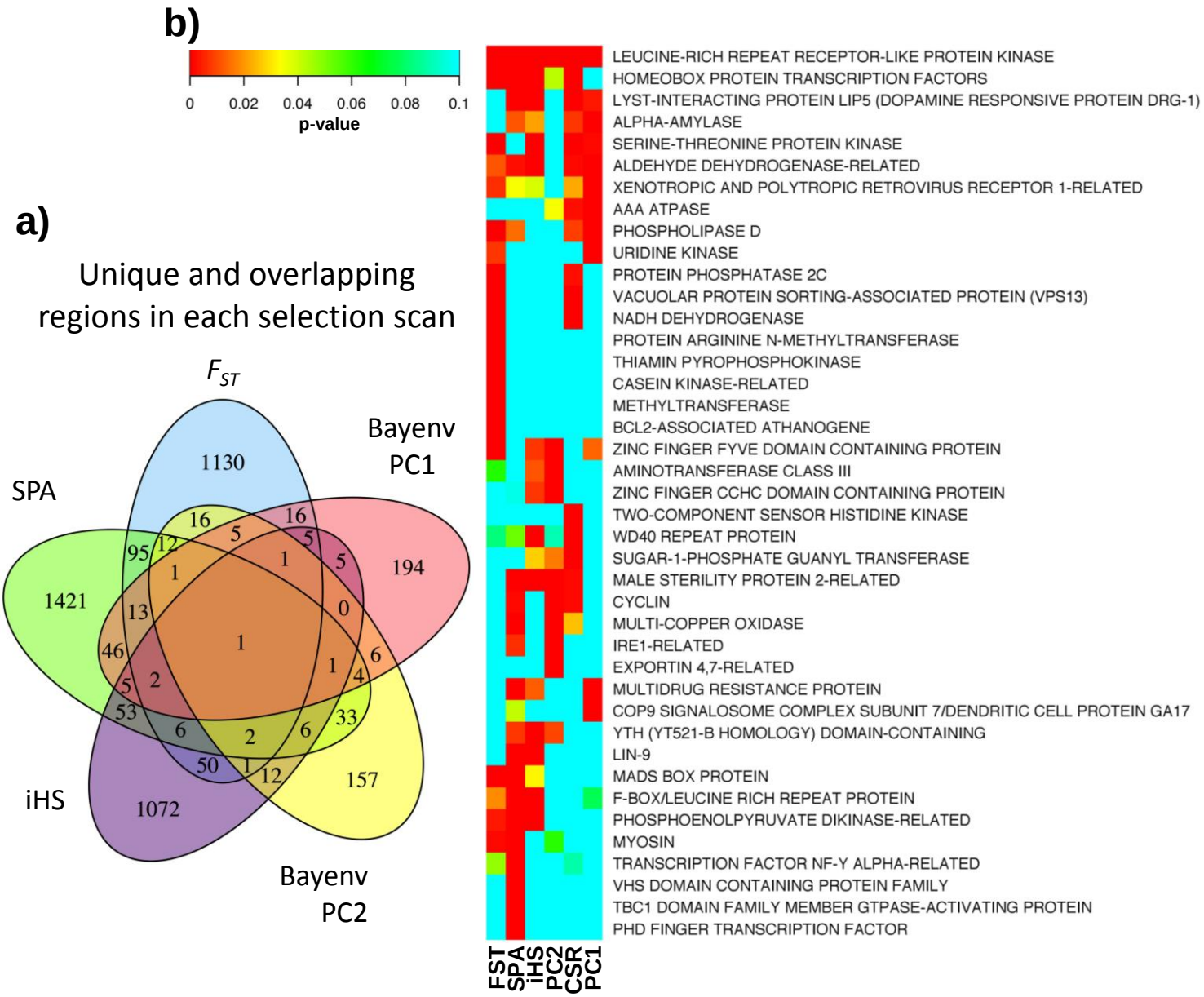

b)
0
0.02
0.04
0.06
0.08
0.1
p-value
a)
Unique and overlapping regions in each selection scan
F_{ST}
Bayenv PC1
SPA
iHS
Bayenv PC2
1130
194
1421
1072
157
95
16
5
12
1
1
5
16
5
13
0
1
46
6
2
4
5
53
6
2
6
33
50
1
12
LEUCINE-RICH REPEAT RECEPTOR-LIKE PROTEIN KINASE
HOMEOBOX PROTEIN TRANSCRIPTION FACTORS
LYST-INTERACTING PROTEIN LIP5 (DOPAMINE RESPONSIVE PROTEIN DRG-1)
ALPHA-AMYLASE
SERINE-THREONINE PROTEIN KINASE
ALDEHYDE DEHYDROGENASE-RELATED
XENOTROPIC AND POLYTROPIC RETROVIRUS RECEPTOR 1-RELATED
AAA ATPASE
PHOSPHOLIPASE D
URIDINE KINASE
PROTEIN PHOSPHATASE 2C
VACUOLAR PROTEIN SORTING-ASSOCIATED PROTEIN (VPS13)
NADH DEHYDROGENASE
PROTEIN ARGININE N-METHYLTRANSFERASE
THIAMIN PYROPHOSPHOKINASE
CASEIN KINASE-RELATED
METHYLTRANSFERASE
BCL2-ASSOCIATED ATHANOGENE
ZINC FINGER FYVE DOMAIN CONTAINING PROTEIN
AMINOTRANSFERASE CLASS III
ZINC FINGER CCHC DOMAIN CONTAINING PROTEIN
TWO-COMPONENT SENSOR HISTIDINE KINASE
WD40 REPEAT PROTEIN
SUGAR-1-PHOSPHATE GUANYL TRANSFERASE
MALE STERILITY PROTEIN 2-RELATED
CYCLIN
MULTI-COPPER OXIDASE
IRE1-RELATED
EXPORTIN 4,7-RELATED
MULTIDRUG RESISTANCE PROTEIN
COP9 SIGNALOSOME COMPLEX SUBUNIT 7/DENDRITIC CELL PROTEIN GA17
YTH (YT521-B HOMOLOGY) DOMAIN-CONTAINING
LIN-9
MADS BOX PROTEIN
F-BOX/LEUCINE RICH REPEAT PROTEIN
PHOSPHOENOLPYRUVATE DIKINASE-RELATED
MYOSIN
TRANSCRIPTION FACTOR NF-Y ALPHA-RELATED
VHS DOMAIN CONTAINING PROTEIN FAMILY
TBC1 DOMAIN FAMILY MEMBER GTPASE-ACTIVATING PROTEIN
PHD FINGER TRANSCRIPTION FACTOR
FST
SPA
iHS
PC2
CSR
PC1

Figure 4.

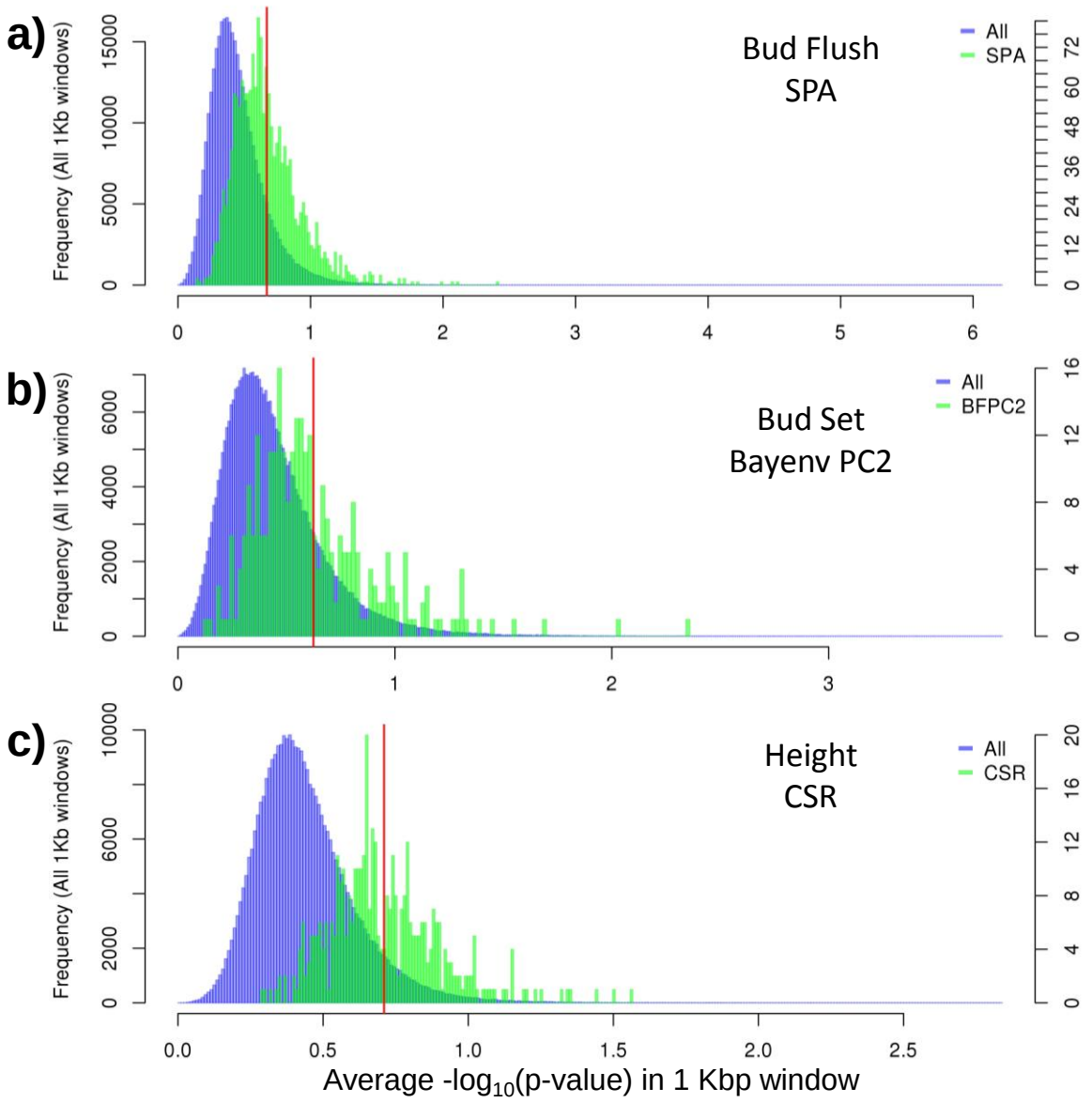

a)
Bud Flush
SPA
All
SPA
Frequency (All 1Kb windows)
b)
Bud Set
Bayenv PC2
All
BFPC2
Frequency (All 1Kb windows)
c)
Height
CSR
All
CSR
Frequency (All 1Kb windows)
Average -log10(p-value) in 1 Kbp window

Figure 5.

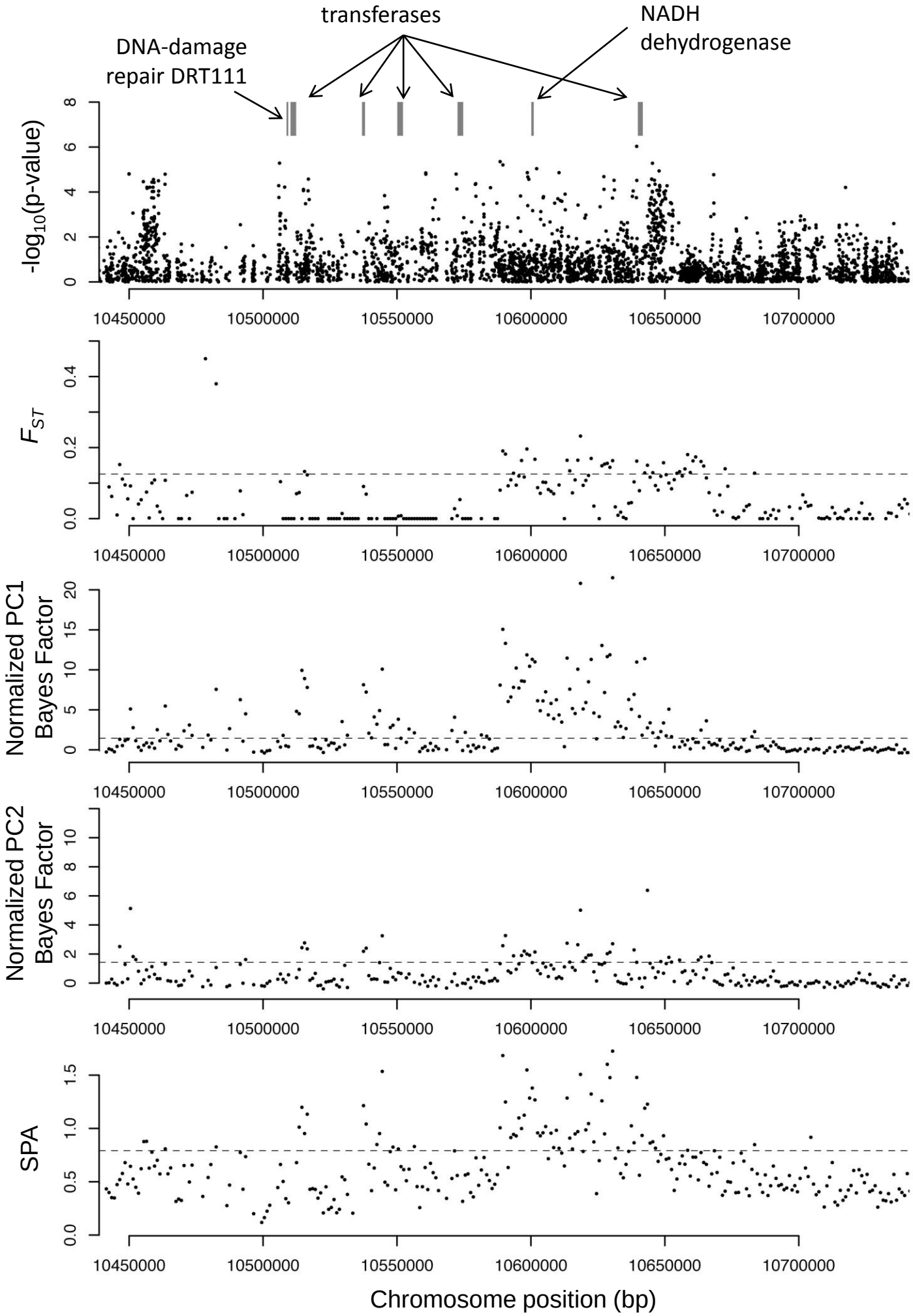
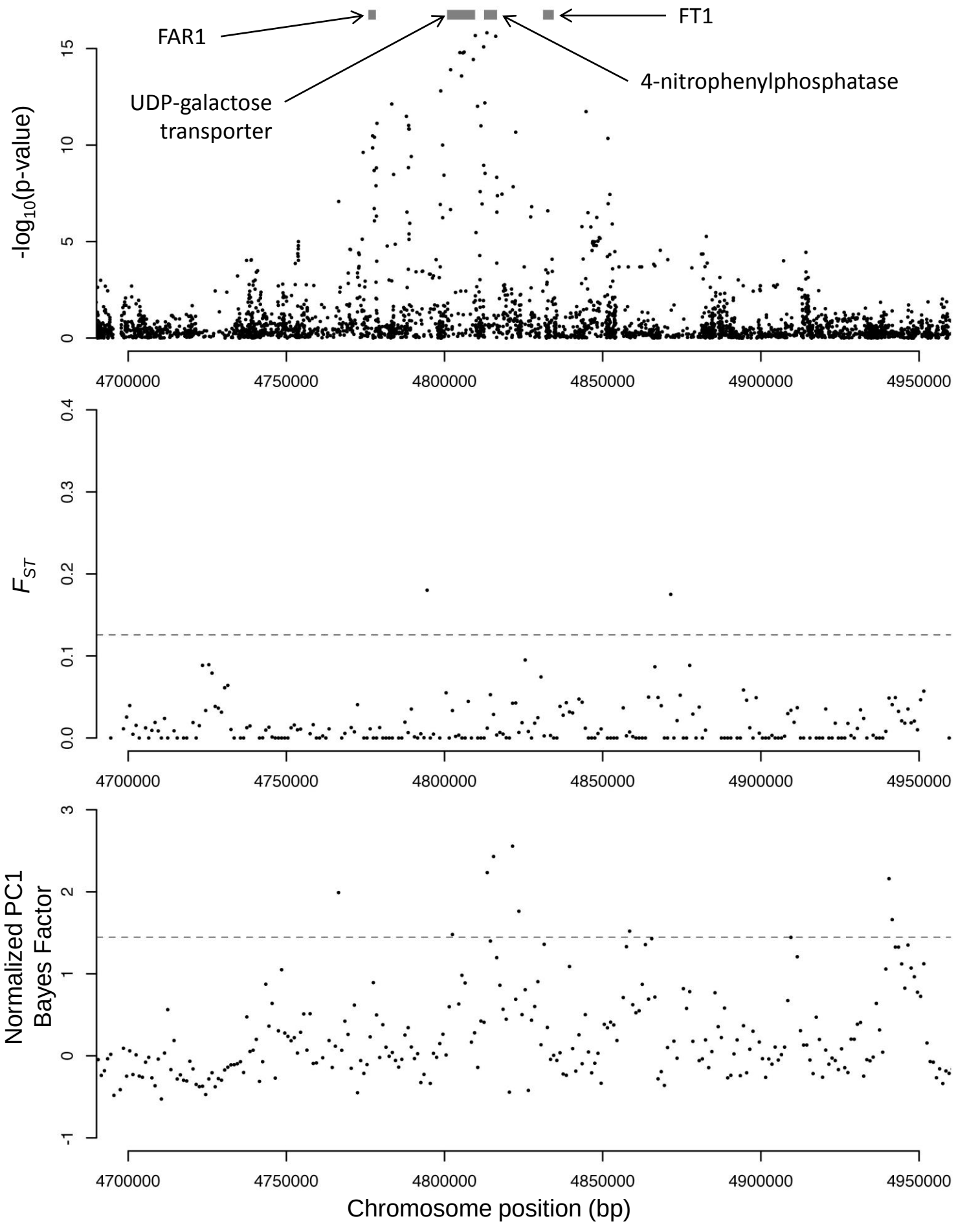

transferases
NADH dehydrogenase
DNA-damage repair DRT111
-log10(p-value)
FST
Normalized PC1 Bayes Factor
Normalized PC2 Bayes Factor
SPA
10450000
10500000
10550000
10600000
10650000
10700000
Chromosome position (bp)

Figure 6.


FAR1
FT1
UDP-galactose transporter
4-nitrophenylphosphatase
-log10(p-value)
0
5
10
15
4700000
4750000
4800000
4850000
4900000
4950000
FST
0.0
0.1
0.2
0.3
0.4
Normalized PC1 Bayes Factor
-1
0
1
2
3
Chromosome position (bp)